AGML, Sommersemester 2020 Vorlesung 7 (6 Seiten)

Lineare Klassifikation II

Anne Driemel Letzte Aktualisierung: 11. Mai 2020

In der letzten Vorlesung haben wir die VC-dimension von Halbräumen analysiert. Die entsprechende Hypothesenklasse $\mathcal{H}$ ist definiert als die Menge von Funktionen der Form $h_{w,u} : \mathbb{R}^d \to \{-1, +1\}$ mit $w \in \mathbb{R}^d, u \in \mathbb{R}$ und

$$h_{w,u}(x) = \begin{cases} +1 \text{ falls } \langle w, x\rangle \geq u \\ -1 \text{ sonst} \end{cases}$$

Lernalgorithmen, die unter Annahme dieser Hypothesenklasse arbeiten, werden unter dem Begriff der linearen Klassifikation zusammengefasst.

Anhand der VC-dimension können wir feststellen, dass eine Hypothesenklasse PAC-lernbar ist. Ein anderer Aspekt ist die Berechnungskomplexität des Lernproblems. Zur Erinnerung, eine Hypothesenklasse ist *effizient* PAC-lernbar, wenn sie mithilfe eines Polynomialzeitalgorithmus $\mathcal{A}$ PAC-lernbar ist.

Wir widmen uns heute der Berechnungskomplexität der linearen Klassifikation. Sei $S = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$ eine beschriftete Trainingsmenge mit $x^{(i)} = (x_1^{(i)}, \ldots, x_d^{(i)}) \in \mathbb{R}^d$ und $y^{(i)} \in \{-1, +1\}$. Die Aufgabe des Lernalgorithmus ist es, Werte für $w \in \mathbb{R}^d$ und $u \in \mathbb{R}$ zu finden sodass der Trainingsfehler

$$\frac{1}{m} \left| \left\{ \; i \in \{1, \ldots, m\} \;\middle|\; h_{w,u}(x^{(i)}) \neq y^{(i)} \; \right\} \right|$$

minimiert wird.

1 Realisierbarer Fall

Im realisierbaren Fall gehen wir davon aus, dass eine Hypothese mit Trainingsfehler 0 existiert. Das entspricht dem Fall, dass die positive und die negative Menge durch eine Hyperebene separierbar sind. In diesem Fall behaupten wir, dass eine solche Hypothese mithilfe linearer Programmierung gefunden werden kann.

Ein lineares Programm bekommt als Eingabe eine Matrix $A \in \mathbb{R}^{m \times n}$ und Spaltenvektoren $b \in \mathbb{R}^m$ und $c \in \mathbb{R}^n$. Die Aufgabe ist es, einen Spaltenvektor $v \in \mathbb{R}^n$ mit $Av \geq b$ zu finden, der $\langle c, v\rangle$ maximiert. Falls dies nicht möglich ist, dann gibt es zwei Möglichkeiten. Entweder existiert kein $v \in \mathbb{R}^n$ welches $Av \geq b$ erfüllt, oder es existiert kein Maximum für $\langle c, v\rangle$ in der Menge der $v \in \mathbb{R}^d$, die $Av \geq b$ erfüllen. Ein lineares Programm kann in polynomieller Zeit in n, m und der Größe der Koordinaten in A, b, c gelöst werden.

Satz 7.1. *Im realisierbaren Fall können wir in polynomieller Zeit in m, d und der Größe der Koordinaten eine Hypothese $h_{\hat{w},\hat{u}} \in \mathcal{H}$ finden, die S korrekt klassifiziert (d.h. $h_{\hat{w},\hat{u}}(x^i) = y^{(i)}$ für alle i).*

Beweis. Wir können die Bedingung $h_{\hat{w},\hat{u}}(x^{(i)}) = y^{(i)}$ wie folgt ausschreiben. Gesucht sind $\hat{w} \in \mathbb{R}^d$ und $\hat{u} \in \mathbb{R}$, sodass für alle $1 \leq i \leq m$ gilt:

(i) $\left\langle \hat{w}, x^{(i)} \right\rangle \geq \hat{u}$ wenn $y^{(i)} = +1$, und

(ii) $\left\langle \hat{w}, x^{(i)} \right\rangle < \hat{u}$ wenn $y^{(i)} = -1$

Wir wollen nun schrittweise ein lineares Programm herleiten, um Werte für $\hat{w}$ und $\hat{u}$ zu finden, die (i) und (ii) erfüllt. Laut der Annahme im Satz existieren w und u, welche diese Bedingungen für $w = \hat{w}$ und $u = \hat{u}$ erfüllen. Daraus folgt

$$\max_{\substack{1 \leq i \leq m \\ y^{(i)} = -1}} \left\langle w, x^{(i)} \right\rangle < u \leq \min_{\substack{1 \leq i \leq m \\ y^{(i)} = +1}} \left\langle w, x^{(i)} \right\rangle \tag{1}$$

wobei w und u unbekannt sind. Da das Maximum auf der linken Seite über eine endliche Menge gebildet wird, existiert ein $u' \in \mathbb{R}$ mit

$$\max_{\substack{1 \leq i \leq m \\ y^{(i)} = -1}} \left\langle w, x^{(i)} \right\rangle < u' < u \leq \min_{\substack{1 \leq i \leq m \\ y^{(i)} = +1}} \left\langle w, x^{(i)} \right\rangle$$

Also gilt für alle $1 \leq i \leq m$, dass

$$y^{(i)} \left\langle w, x^{(i)} \right\rangle > y^{(i)} u'$$

Weiter können wir die rechte Seite subtrahieren und bekommen

$$y^{(i)} \left\langle w, x^{(i)} \right\rangle - y^{(i)} u' > 0$$

Es folgt, dass ein Wert $\gamma > 0$ existiert, sodass für alle $1 \leq i \leq m$

$$y^{(i)} \left\langle w, x^{(i)} \right\rangle - y^{(i)} u' \geq \gamma$$

Das können wir äquivalent umformen zu

$$\left\langle y^{(i)} x^{(i)}, w'' \right\rangle - y^{(i)} u'' \geq 1 \tag{2}$$

mit $w'' = \frac{w}{\gamma}$ und $u'' = \frac{u'}{\gamma}$.

Wir können nun die Zeilen der Matrix A des linearen Programms definieren als $(d+1)$-dimensionale Zeilenvektoren

$$a_i = (y^{(i)} x_1^{(i)}, y^{(i)} x_2^{(i)}, \ldots, y^{(i)} x_d^{(i)}, -y^{(i)})$$

für $1 \leq i \leq m$. Für b wählen wir den m-dimensionaler Spaltenvektor $(1, \ldots, 1)$ und für c den m-dimensionalen Spaltenvektor $(0, \ldots, 0)$.

Das lineare Programm findet dann ein $v = (v_1, \ldots, v_n)$ mit $Av \geq b$, sodass $\langle c, v \rangle$ maximiert wird. Dabei ist $\langle c, v \rangle = 0$ für alle $v \in \mathbb{R}^n$ und wir interessieren uns eigentlich nur für den ersten Teil der Bedingung.

Laut unserem linearen Programm haben wir dann ein v, das (2) erfüllt mit $v = (w_1'', \ldots, w_d'', u'')$. Durch unsere Herleitung aus w und u wissen wir, dass solch ein v existieren muss. Das heisst, wir können nun $w'' \in \mathbb{R}^n$ und u'' aus den Koordinaten von v ablesen. Wir wählen nun

$$\hat{w} = \frac{w''}{\|w''\|}$$

und

$$\hat{u} = \min_{\substack{1 \leq i \leq m \\ y^{(i)} = +1}} \left\langle \hat{w}, x^{(i)} \right\rangle$$

und geben diese zurück als Lösung. Tatsächlich klassifiziert die Hypothese $h_{\hat{w},\hat{u}}$ alle Punkte in S korrekt, da

$$\hat{w} = \frac{w''}{\|w''\|} = \frac{(\frac{w_1}{\gamma}, \ldots, \frac{w_d}{\gamma})}{\|(\frac{w_1}{\gamma}, \ldots, \frac{w_d}{\gamma})\|} = \frac{\frac{1}{\gamma}w}{\frac{1}{\gamma}\|w\|} = \frac{w}{\|w\|}$$

und weil aus (1) folgt, dass auch

$$\max_{\substack{1\leq i\leq m \\ y^{(i)}=-1}} \left\langle \frac{w}{\|w\|}, x^{(i)} \right\rangle < \min_{\substack{1\leq i\leq m \\ y^{(i)}=+1}} \left\langle \frac{w}{\|w\|}, x^{(i)} \right\rangle$$

gilt. □

2 Nicht-Realisierbarer Fall

Im nicht-realisierbaren Fall gehen wir *nicht* davon aus, dass die positive Menge und die negative Menge durch eine Hyperebene separierbar sind. In diesem Fall ist es NP-schwer einen Halbraum zu finden, der den Trainingsfehler minimiert. Wir zeigen dies im speziellen Fall der Hypothesenklasse $\mathcal{H}_0$ von Funktionen der Form $h_w : \mathbb{R}^d \to \{-1, +1\}$ mit $w \in \mathbb{R}^d$ und

$$h_w(x) = \begin{cases} +1 \text{ falls } \langle w, x\rangle \geq 0 \\ -1 \text{ sonst} \end{cases}$$

In der letzten Vorlesung hatten wir gesehen, dass diese Klasse, mithilfe einer Transformation in einen höherdimensionalen Raum, auch allgemeine lineare Klassifikatoren darstellen kann.

Wir zeigen die NP-Schwerheit des Lernproblems unter $\mathcal{H}_0$ mithilfe einer Reduktion von dem folgenden NP-schweren Problem.

Definition 7.2 (MAX-E2-SAT)**.** *Gegeben eine Menge von m Klauseln über n booleschen Variablen $x_1, \ldots, x_n$, wobei jede Klausel genau zwei Literale (negierte oder nicht-negierte Variablen) enthält. Finde eine Wahrheitsbelegung der Variablen, welche die Anzahl der erfüllten Klauseln maximiert.*

Beispiel 7.3. *Sei $\{(x_1 \vee x_2), (\overline{x_1} \vee \overline{x_2}), (\overline{x_2} \vee \overline{x_3}), (\overline{x_1} \vee x_3)\}$ eine Menge von Klauseln. Eine Wahrheitsbelegung, welche die Anzahl der erfüllten Klauseln maximiert, ist $x_1 = 1$, $x_2 = 0$, $x_3 = 1$. Diese Wahrheitsbelegung ist maximal, da alle Klauseln durch sie erfüllt werden.*

Satz 7.4 (Håstad)**.** *Falls $P \neq NP$, dann existiert kein polynomieller Algorithmus für MAX-E2-SAT. (ohne Beweis)*

Wir wollen aus dem Satz von Håstad folgern, dass auch das Lernproblem über $\mathcal{H}_0$ NP-schwer ist. Das heisst, wir wollen den folgenden Satz zeigen.

Satz 7.5. *Falls $P \neq NP$, dann existiert kein polynomieller Algorithmus, der ein $h \in \mathcal{H}_0$ findet welches den Trainingsfehler minimiert.*

Gegeben sei eine Menge $\mathcal{I}$ von m Klauseln über n Variablen $x_1, \ldots, x_n$ als Eingabe für das MAX-E2-SAT Problem. Wir transformieren diese Eingabe in eine Eingabe $\mathcal{I}'$ für das Lernproblem über $\mathcal{H}_0$. Wir definieren für jede Klausel C einen Punkt $\phi(C) \in \mathbb{R}^n$ mithilfe einer Funktion

$$\phi_j(C) = \begin{cases} 1 \text{ falls } x_j \in C \\ -1 \text{ falls } \overline{x_j} \in C \\ 0 \text{ sonst} \end{cases}$$

Sei $\phi(C) = (\phi_1(C), \ldots, \phi_n(C))$. Wir geben diesem Punkt ein positives Label.

Zusätzlich definieren wir für jede Klausel C über Variablen x_i, x_j eine Menge von vier Punkten $\{e_i, e_j, -e_i, -e_j\}$, wobei e_i den Einheitsvektor von $\mathbb{R}^n$ bezeichnet, der überall 0 ist, und nur an der $i$ten Koordinate eine 1 hat. Wir geben diesen Punkten ein negatives Label und fügen sie in zweifacher Ausführung hinzu. Die Klausel C erzeugt also eine beschriftete Menge

$$\Phi(C) = \big\{(\phi(C), +1), (e_i, -1), (e_i, -1), (e_j, -1), (e_j, -1), (-e_i, -1), (-e_i, -1), (-e_j, -1), (-e_j, -1)\big\}$$

Die Eingabe $\mathcal{I}'$ für das Lernproblem besteht nun aus der Vereinigung dieser beschrifteten Punktmengen über alle Klauseln. Beachte, dass in der erzeugten Menge Punkte mehrfach vorkommen.

Definition 7.6. *Sei $h_w \in \mathcal{H}_0$ eine Hypothese mit $w = (w_1, \ldots, w_n)$. Wir definieren eine Funktion $\alpha : \mathbb{R}^n \to \{0,1\}^n$ mit*

$$\alpha_i(w) = \begin{cases} 1 \text{ falls } w_i \geq 0 \\ 0 \text{ sonst} \end{cases}$$

als $\alpha(w) = (\alpha_1(w), \ldots, \alpha_n(w))$. Die Funktion bildet die Hypothese h_w auf eine Wahrheitsbelegung für die Variablen $x_1, \ldots, x_n$ ab, indem wir $x_i = \alpha_i(w)$ setzen.

Sei $h_w \in \mathcal{H}_0$ eine Hypothese, die den Trainingsfehler auf Eingabe $\mathcal{I}'$ minimiert. Wir behaupten, dass $\alpha(w)$ die Anzahl der erfüllten Klauseln in $\mathcal{I}$ maximiert. Um das zu zeigen, müssen wir zunächst ein paar strukturelle Eigenschaften unserer Konstruktion zeigen.

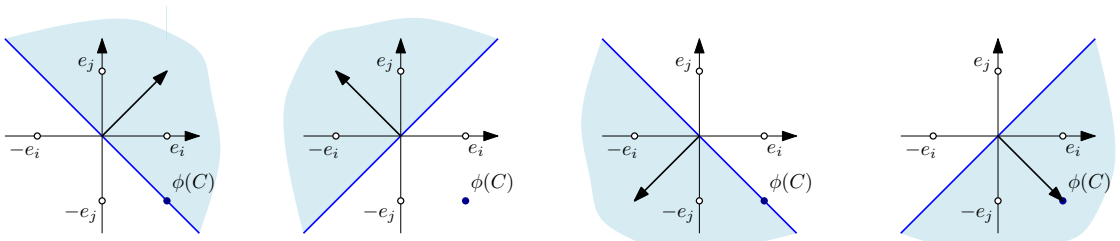
Behauptung 7.7. *Wenn für ein $k \geq 0$ eine Wahrheitsbelegung $a \in \{0,1\}^n$ existiert, die k Klauseln von $\mathcal{I}$ erfüllt, dann existiert ein $h_w \in \mathcal{H}_0$, welches $k + 4m$ Punkte in $\mathcal{I}'$ korrekt klassifiziert.*

Beweis. Dafür setzen wir

$$w_i = \begin{cases} 1 \text{ falls } a_i = 1 \\ -1 \text{ falls } a_i = 0 \end{cases}$$

Dann ist $\langle w, \phi(C) \rangle \geq 0$ genau dann wenn die Wahrheitsbelegung a die Klausel C erfüllt. Das lässt sich leicht durch eine Fallanalyse zeigen, die wir hier nicht ausführen. Ferner werden genau 4 negative Punkte von $\Phi(C)$ korrekt klassifiziert. Damit ist Behauptung 7.7 bewiesen. □

Beispiel 7.8. *Sei $C = (x_i \vee \overline{x_j})$, dann ist $\phi_i(C) = 1$ und $\phi_j(C) = -1$ und alle anderen Koordinaten von $\phi(C)$ sind gleich null. Das heisst, $\phi(C)$ liegt in dem linearen Unterraum, der durch die Einheitsvektoren e_i und e_j aufgespannt wird. Daher können wir uns die vier Hypothesen aus obigem Beweis, die den vier Wahrheitsbelegungen von x_i und x_j entsprechen, wie folgt vorstellen:*



(a) $x_i = 1$, $x_j = 1$
$(x_i \vee \overline{x_j}) = 1$

(b) $x_i = 0$, $x_j = 1$
$(x_i \vee \overline{x_j}) = 0$

(c) $x_i = 0$, $x_j = 0$
$(x_i \vee \overline{x_j}) = 1$

(d) $x_i = 1$, $x_j = 0$
$(x_i \vee \overline{x_j}) = 1$

Der Fall (b) ist die einzige Belegung, wo die Klausel nicht erfüllt ist. Das ist auch der einzige Fall, in dem $\phi(C)$ nicht korrekt klassifiziert wird. Weiter ist die Anzahl der negativen Punkte, die von h_w als negativ klassifiziert werden, immer genau $4m$. Also werden genau $k+4m$ Punkte korrekt klassifiziert. Die anderen Klauseln können auf die gleiche Art analysiert werden.

Behauptung 7.9. *Sei $h_w \in \mathcal{H}_0$ mit $w = (w_1, \ldots, w_n) \in \mathbb{R}^n$ eine Hypothese, die den Trainingsfehler minimiert, dann ist $w_i \neq 0$ für alle $1 \leq i \leq n$.*

Beweis. Sei $w_i = 0$ für eine Hypothese h_w. Sei C eine Klausel über Variablen x_i und x_j. Dann ist $\langle w, e_i \rangle \geq 0$, sowie $\langle w, -e_i \rangle \geq 0$. Gleichzeitig ist entweder $\langle w, e_j \rangle \geq 0$, oder $\langle w, -e_j \rangle \geq 0$. Da diese Punkte in zweifacher Ausführung in $\Phi(C)$ vorkommen, klassifiziert h_w also mindestens 6 Punkte von $\Phi(C)$ falsch, also höchstens 3 Punkte korrekt. Gleichzeitig klassifiziert $h_{w'}$ mit einem beliebigen $w' = (w'_1, \ldots, w'_n)$ mit $w'_j \neq 0$ für alle $1 \leq j \leq n$ mindestens 4 negative Punkte pro Klausel korrekt. Damit ist Behauptung 7.9 bewiesen. □

Behauptung 7.10. *Sei $h_w \in \mathcal{H}_0$ mit $w \in \mathbb{R}^n$ eine Hypothese, die den Trainingsfehler minimiert. Sei $\phi(C)$ ein Punkt, der durch h_w korrekt klassifiziert wird, dann wird die Klausel C durch $\alpha(w)$ erfüllt.*

Beweis. Das kann wieder durch eine Fallanalyse gezeigt werden. Sei C die Klausel $(x_i \vee x_j)$. Dann ist $\phi_i(C) = 1$ und $\phi_j(C) = 1$ und alle anderen Koordinaten sind gleich null. Daher gilt für alle $w \in \mathbb{R}^n$

$$\langle w, \phi(C) \rangle \geq 0 \quad \Leftrightarrow \quad w_i + w_j \geq 0$$

Wir unterscheiden die folgenden Fälle.

(a) $(w_i > 0, w_j > 0) \Rightarrow (x_i = 1, x_j = 1) \Rightarrow C$ ist durch $\alpha(w)$ erfüllt

(b) $(w_i > 0, w_j < 0) \Rightarrow (x_i = 1, x_j = 0) \Rightarrow C$ ist durch $\alpha(w)$ erfüllt

(c) $(w_i < 0, w_j > 0) \Rightarrow (x_i = 0, x_j = 1) \Rightarrow C$ ist durch $\alpha(w)$ erfüllt

(d) $(w_i < 0, w_j < 0) \Rightarrow (w_i + w_j < 0) \Rightarrow \phi(C)$ wird nicht korrekt klassifiziert

Wir können annehmen, dass $w_i \neq 0$ und $w_j \neq 0$, da sonst h_w nicht optimal ist (Behauptung 7.9). Somit ist die obige Fallanalyse für die betrachtete Klausel C vollständig. Die anderen Möglichkeiten für C sind die Klauseln $(x_i \vee \overline{x_j}), (\overline{x_i} \vee x_j), (\overline{x_i} \vee \overline{x_j})$. In diesen Fällen kann die Behauptung analog gezeigt werden, was wir hier nicht ausführen. Damit wäre Behauptung 7.10 bewiesen. □

Beweis von Satz 7.5. Wir können nun alles zusammenführen und unseren Satz beweisen. Laut Behauptung 7.7 existiert für jede Wahrheitsbelegung mit k erfüllten Klauseln von $\mathcal{I}$ eine Hypothese, die $k + 4m$ Punkte in $\mathcal{I}'$ korrekt klassifiziert. Gleichzeitig folgt aus Behauptung 7.9 für jedes h_w, das den Trainingsfehler auf $\mathcal{I}'$ minimiert, dass die Anzahl der negativen Punkte, die durch h_w korrekt klassifiziert werden, gleich $4m$ ist. Wenn h_w also $k + 4m$ Punkte korrekt klassifiziert, dann sind k Punkte davon positiv. Aus Behauptung 7.10 folgt dann, dass h_w eine Wahrheitsbelegung $\alpha(w)$ impliziert, die mindestens k Klauseln von $\mathcal{I}$ erfüllt. Wenn es also eine Wahrheitsbelegung gibt, die k Klauseln in $\mathcal{I}$ erfüllt, dann gibt unsere Reduktion mithilfe eines Lernalgorithmus für $\mathcal{I}'$ eine Wahrheitsbelegung zurück, die mindestens k Klauseln in $\mathcal{I}$ erfüllt. Gäbe es also einen polynomiellen Algorithmus für das Lernproblem, dann gäbe es auch einen polynomiellen Algorithmus für MAX-E2-SAT. Damit folgt Satz 7.5 aus Satz 7.4. □

Referenzen

- Foundations of Machine Learning, Kapitel 5.2.
- Understanding Machine Learning, Kapitel 9.1.1.
- Bernhard Korte und Jens Vygen, Combinatorial Optimization–Theory and Algortihms, Third Edition, Springer.
- Shai Ben-David , Nadav Eiron , Philip M. Long, "On the Difficulty of Approximately Maximizing Agreements", Journal of Computer and System Sciences, 2000.